Análisis de estrategias para clasificar contenidos en foros de discusión

Valeria Zoratto, Nadina Martinez Carod, Facundo Otermin, Gabriela Aranda

{vzoratto|nadina.martinez}@fi.uncoma.edu.ar

Abstract. La información contenida en los foros de discusión de la Web es, en la mayoría de los casos, considerada muy valiosa por usuarios con problemas o necesidades similares. Por ese motivo, en el último tiempo se han multiplicado los esfuerzos para recuperar, analizar y reutilizar la información que se extrae de dichos hilos de discusión. En trabajos previos se ha presentado una estrategia de análisis de hilos de discusión basada en la herramienta de búsqueda Lucene. Siguiendo esa línea, en este artículo se presenta una extensión de la metodología anterior en la que se combina la funcionalidad de la base de datos léxica WordNet y el parser de lenguaje natural Standford, con el objetivo de evaluar la inclusión de sinónimos considerando la estructura gramatical de los hilos bajo estudio.

1 Introducción

Con la evolución de la Web, han aparecido distintas plataformas de trabajo colaborativo que permiten que los usuarios se comuniquen y trabajen en forma conjunta sin importar que estén reunidos en un mismo lugar físico ni que lo hagan en el mismo instante. Estas plataformas tienen como objetivo general, además de compartir información en distintos formatos, permitir el intercambio de opiniones y conocimiento. Ejemplos de estas plataformas, ampliamente utilizadas en la actualidad, son las Wikis, los Weblogs y los foros de discusión. Estos últimos son especialmente interesantes dado que habitualmente son utilizados para, ante una dificultad, solicitar ayuda a usuarios más expertos en dominios específicos [1]. En particular, los foros de discusión sobre temáticas relacionadas al desarrollo y mantenimiento de software, tienen un gran volumen de contenido útil, producido por sus usuarios, que es deseable y valioso que pueda ser extraído y reutilizado [2]. Además, los foros tienen una estructura única [3],constituida por un conjunto de hilos, donde cada uno está compuesto por un título, una pregunta principal, y luego una serie de respuestas que representan el debate sobre dicha pregunta.

Usualmente, ante un problema técnico, los usuarios utilizan los motores de búsqueda multi-propósito para acceder a las discusiones de los foros, lo que los lleva a recorrer varias páginas hasta encontrar un problema similar al suyo entre diversos hilos de uno o de varios foros. Sin embargo, a veces la solución encontrada no es la adecuada para dicho problema y es necesario probar varias posibles soluciones hasta encontrar la correcta, transformándose en una tarea que insume gran cantidad de tiempo.

Con el objetivo de minimizar el trabajo manual, en [4] se presentaron las características de una herramienta para facilitar la búsqueda de soluciones a problemas comunes, estableciendo prioridades entre las soluciones disponibles en foros de discusión dispersos en la Web. Una componente de la herramienta propuesta, desarrollada en [5], evalúa la información obtenida de los foros a partir del análisis léxico de sus datos. A fin de mejorar los resultados obtenidos hasta el momento, se propone una extensión de dicha herramienta que permite evaluar la inclusión de sinónimos en la clasificación de los hilos de discusión, empleando WordNet (base de datos léxica ampliamente utilizada para el idioma inglés) [6], y Stanford Parser[1], para analizar la estructura gramatical del hilo.

El proceso que se busca mejorar está representado en la Figura 1, donde se muestra el desarrollo de la clasificación de hilos de acuerdo a un conjunto de entidades reconocibles, llamadas *documentos de referencia.* Como este ejemplo se basa en el lenguaje de programación Java, los documentos de referencia forman parte del repositorio de especificación de clases de Oracle (versión 5) [2], sin embargo, dicho conjunto de documentos puede variar al enfocar el estudio en otras temáticas.

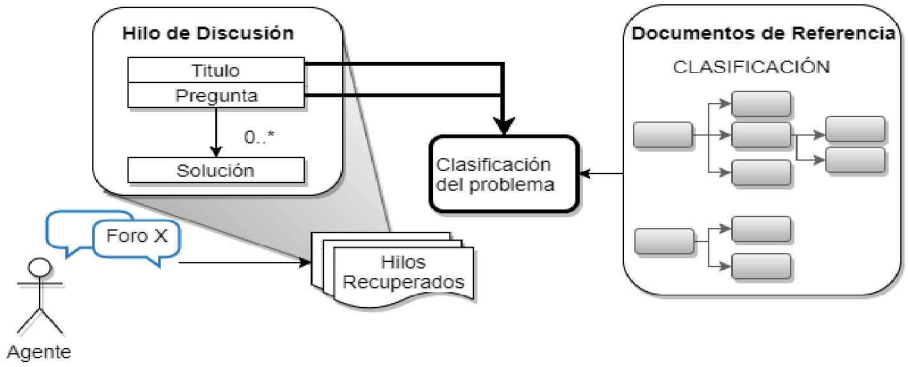


Fig. 1. Proceso de clasificación de hilos de acuerdo a documentos de referencia

Teniendo en cuenta este objetivo, en la Sección 2 se detallan los antecedentes que llevaron a la realización del primer caso de estudio. Posteriormente, en la Sección 3 se detalla la modificación de la herramienta propuesta en [5] y en la Sección 4 se evalúan los resultados alcanzados con dicha modificación. Luego, en la Sección 5 se comparan los resultados originales con los obtenidos en la versión modificada. Finalmente, en la Sección 6 se presentan las conclusiones y trabajo futuro.

[1] https://nlp.stanford.edu/software/tagger.shtml
[2] http://docs.oracle.com/javase/1.5.0/docs/

2 Antecedentes

En un foro de discusión, cada hilo está compuesto por un título y una pregunta principal, y luego se encadenan una secuencia de respuestas que representan el debate sobre dicha pregunta. Por el otro lado, la estructura de un documento Oracle tiene el nombre de la clase (*Class Name* en el documento Oracle), una lista de las interfaces que implementa (*Implemented Interfaces*), las subclases conocidas (*Known Subclasses*), y luego se describen los métodos de la clase con sus parámetros, comentarios y en algunos casos ejemplos de utilización. Estas estructuras se muestran esuqmatizadas en la Figura 2.

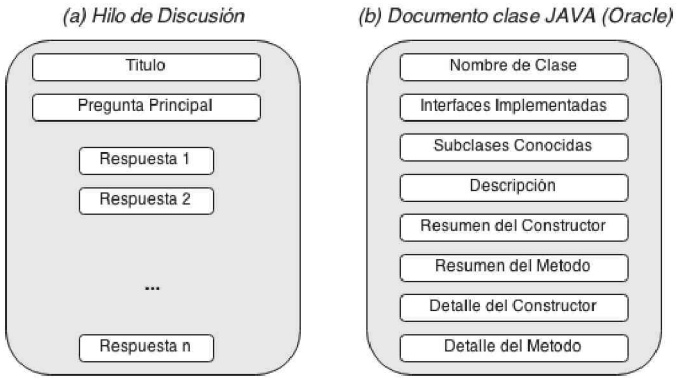


Fig. 2. Estructuras de los documentos estudiados

En base a la información contenida en las secciones presentes en cada tipo de documento (hilos de discusión y documentos Oracle), y a la tarea objeto de esta investigación de clasificar los hilos de foros de discusión de la manera más apropiada, se definieron las siguientes hipótesis:

- Hipótesis A: Utilizar mayor cantidad de información sobre cada una de las clases Java documentadas en Oracle permite clasificar los hilos de discusión relacionados a ellas de forma más precisa.
- Hipótesis B: Utilizar más información sobre el problema explicado en los hilos de discusión permite clasificarlos de forma más precisa respecto a los documentos Oracle de las clases Java.

A efectos de evaluar dichas hipótesis, en [5] se propuso una metodología de clasificación de hilos reales de discusión técnicos. Para evaluar las hipótesis propuestas, se establecieron las siguientes fases:

- Fase 1. Recuperación de documentos: Recuperar los hilos de discusión del foro de discusión y las especificaciones de clases Java del repositorio de Oracle (documentaos de referencia).
- Fase 2. Clasificación por expertos: Los hilos de discusión recuperados en la fase 1 son analizados por tres expertos para identificar las clases Java más relacionadas con cada uno.

- Fase 3. Pre-procesamiento de documentos: Preparar los documentos descargados para su posterior análisis, eliminando código html irrelevante e información innecesaria,
- Fase 4. Indexación de documentos de referencia: Utilizar la API Lucene [7] para indexar los documentos de referencia.
- Fase 5. Búsqueda de documentos relevantes: Utilizar la API Lucene para determinar la relación entre cada hilo de discusión y los documentos Oracle de las clases Java.
- Fase 6: Evaluación: Contrastar los resultados de los expertos con los retornados por la herramienta.

Dado que la herramienta Lucene permite redefinir el conjunto de palabras *stop-words*[3], se modificó el conjunto predefinido para que no se consideren como tal las palabras reservadas del lenguaje Java (for, then, if, this) y se agregaron otras que han sido consideradas poco representativas en dicho contexto.

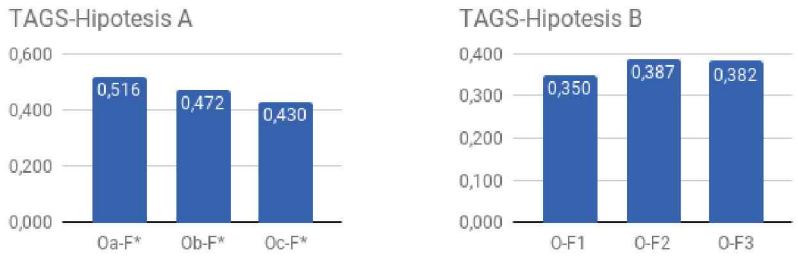
Esta metodología fue aplicada sobre un conjunto de hilos de discusión referidos a problemas sobre el uso del lenguaje Java. En particular, se recuperaron 50 hilos del foro de discusión Stack Overflow[4], utilizando la funcionalidad de filtro por tags de dicho sitio, seleccionando el tag *java*. Uno de los problemas detectados a partir de la ejecución de un caso de estudio fue que algunos documentos Oracle aparecían relacionados a la mayoría de los hilos de discusión con un valor alto, y a partir del análisis correspondiente se detectó que se trataba de nombres de clases que representaban vocablos de uso común en el lenguaje natural dentro del ambiente de programación, como por ejemplo: *Class*, *Error*, *Type*, etc., por lo que se reprodujo la fase de indexación, pero en este caso eliminando dichos documentos del conjunto de referencia.

En [8] se presentó un análisis de dos casos de estudio siguiendo la metodología mencionada, en la que se observan los resultados obtenidos diferenciando, por un lado la cantidad y calidad de información obtenida de los hilos (casos F_1, F_2, F_3) y luego considerando diferentes porcentajes de información de la documentación extraída de Oracle (casos O_a, O_b, O_c). Los resultados obtenidos se grafican en la Figura 3 donde se puede observar que en el caso (a), al contar con mayor información de los documentos Oracle (caso O_c) la performance es mas baja que utilizando los documentos que contienen menos información (casos O_ay O_b). En el caso (b), la performance es mayor cuando se consideran los documentos que contienen el título del hilo y la pregunta principal (caso F_2), pero al utilizar el texto completo del hilo (caso F_3) mejora los resultados en comparación de las ocurrencias donde sólo se considera el título del mismo (caso F_1).

A partir de los resultados obtenidos en ese caso de estudio, surgió la inquietud de modificar el proceso incorporando sinónimos al texto recuperado de los hilos de discusión. Enfocado en la hipótesis B que sugiere que la precisión aumenta a medida que se cuenta con más información sobre el problema, se plantea la siguiente hipótesis, que será desarrollada a continuación:

[3] Palabras que carecen de significado por sí solas y no brindan información acerca del contenido del texto

[4] http://stackoverflow.com/



(a) (b)

Fig. 3. Análisis de performance según las hipótesis planteadas

– HIPÓTESIS B.2: Agregar sinónimos al problema explicado en los hilos de discusión permite clasificarlos de forma más precisa respecto a los documentos Oracle de las clases Java.

3 Mejora en el procesamiento de hilos

La incorporación de sinónimos para mejorar los resultados obtenidos al clasificar los hilos de discusión, requiere que se extienda la etapa de procesamiento previa a la fase 5. Surge así una modificación al proceso propuesto en [8], tal como se presenta en la Figura 4. En dicha modificación, se agrega una nueva fase, entre las fases 3 y 5 (explicadas en la Sección 2), para implementar un segundo procesamiento de los hilos obtenidos en la fase 3, que los enriquece incorporando sinónimos, mediante la utilización de WordNet y del parser Stanford.

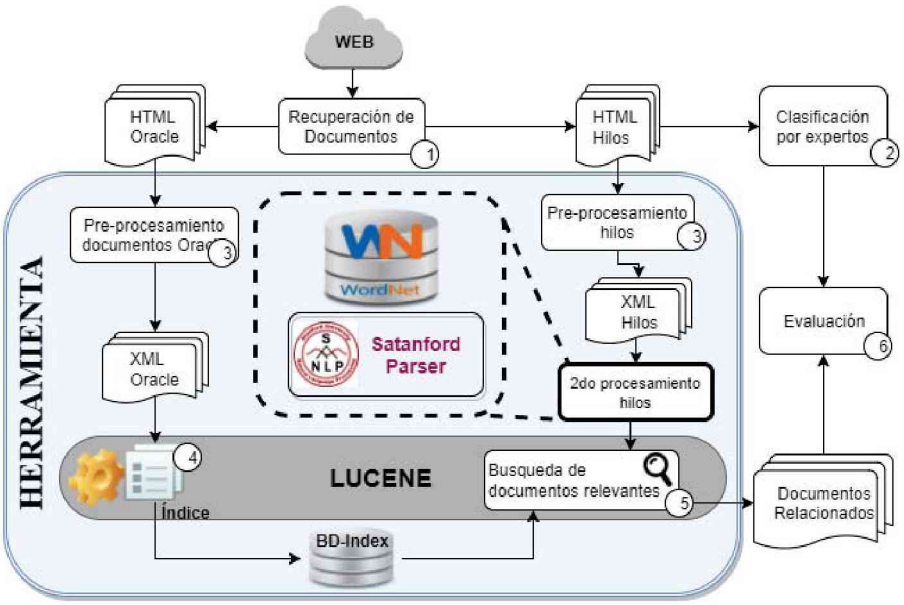


Fig. 4. Fases del desarrollo utilizando WordNet y Stanford parser

3.1 Procesamiento de los hilos para agregar sinónimos

En esta fase se utilizan como entrada los documentos pre-procesados durante la Fase 3 (*pre-procesamiento hilos*), donde los documentos *html* se traducen a formato *xml*, descartando código *html* irrelevante y agregando etiquetas para delimitar los bloques de interés. Además, en la Fase 3 se crean tres versiones de cada documento, tanto para los hilos como para los documentos Oracle, considerando un subconjunto de las partes que componen cada documento, como se presenta en la Tabla 1.

Table 1. Versiones de los diferentes hilos y documentos de referencia

Hilos		**Documentos de referencia**	
F_1	Sólo el título del hilo	O_a	Sólo el nombre de la clase
F_2	Título del hilo y la pregunta principal	O_b	Nombre de la clase y los nombres de todos sus métodos
F_3	Texto completo del hilo (título, pregunta principal, respuestas)	O_c	Todas las secciones del documento (excepto las secciones "Detalles de constructor" y "Detalle de métodos").

En base a las tres versiones de cada tipo de documento, se establecieron nueve combinaciones que formaron la base del análisis, tal como se muestra en la Tabla 2.

Table 2. Combinaciones de los tipos de documentos a analizar

	Sólo título (F_1)	**Título + pregunta (F_2)**	**Texto completo (F_3)**
Sólo el nombre de la clase (O_a)	O_aF_1	O_aF_2	O_aF_3
Nombre de la clase y métodos (O_b)	O_bF_1	O_bF_2	O_bF_3
Texto completo (O_c)	O_cF_1	O_cF_2	O_cF_3

Para considerar los sinónimos en la búsqueda de documentos de referencia relacionados, se utiliza el Stanford Parser para etiquetar las palabras según la gramática de la oración, teniendo en cuenta el contexto en el que se utiliza cada palabra. De esta manera, la palabra se clasifica como verbo, adjetivo, adverbio, sustantivo, u otro. El parser detecta qué grupos de palabras van juntas (como "frases") y qué palabras son el sujeto u objeto de un verbo. Por ejemplo, para la frase "*My dog also likes eating sausage.*", el parser etiqueta la cadena a "My/*PRP$* dog/*NN* also/*RB* likes/*VBZ* eating/*VBG* sausage/*NN* ./.". De la misma manera que [9], en esta instancia del estudio nos interesamos en el análisis de sustantivos y sus modificadores (adjetivos y adverbios).

Una vez que se cuenta con todas las palabras del hilo etiquetadas según la estructura gramatical, se utiliza WordNet para recuperar los sinónimos (según el tipo correspondiente) y se agregan a la cadena de búsqueda teniendo en cuenta que la palabra no sea un stopword. Por otro lado, se han tomado como caso

especial los valores numéricos que suelen agregar sinónimos innecesarios. Por ejemplo, para el valor 7, Wordnet considera sinónimos los términos *seven*, *VII*, *Sevener*, *heptad*, *septet* y *septenary*, que van más allá del significado del número en sí, por eso se decidió ignorar los sinónimos de los números en esta etapa del estudio.

Una vez que se han obtenido las nuevas versiones de los hilos, se procede a ejecutar la Fase 5 para obtener los documentos relevantes para cada consulta.

4 Evaluación y Analisis de resultados

Dado que el objetivo de los sistemas de recuperación de información es tratar de maximizar la cantidad de documentos recuperados que sean relevantes, se utilizará la medida F-Measure para evaluar la performance del método de recuperación, dado que brinda mayor información que las medidas de precisión y *recall*, y que es una de las medidas más utilizadas al evaluar técnicas de recuperación de información [10].

Un enfoque adicional es calcular estas medidas con valores de cortes (*cut-off*) sobre la cantidad de respuestas válidas obtenidas por el recuperador de información utilizado. Es decir, en el caso de corte N se analizan la cantidad de documentos relevantes, no relevantes, precisión, recall y F-Measure que se obtienen al considerar solamente los primeros N documentos adquiridos en el proceso de recuperación.

Es importante evaluar los resultados estimando valores de cortes sobre la cantidad de respuestas validas, es por ello que, teniendo en cuenta lo realizado en [5,8], se utilizaron las mismas medidas de corte, conservando como corte inicial N=2 y siguiendo una escala de N=3, 4 y 5.

Conservando las combinaciones planteadas anteriormente (Sección 3.1) se obtuvieron 9 posibilidades sin utilizar las clases Oracle consideradas stopwords. Las mediciones para los valores de corte mencionados anteriormente, utilizando la medida F-Measure, se presentan en la Tabla 3.

En la Figura 5 se muestran las tendencias de la performance en este caso de estudio. Se puede observar que al considerar el título y la pregunta principal del hilo (F_2) mejora la performance de la clasificación respecto a considerar sólo el título (F_1) o el hilo completo (F_3). En los casos en que la búsqueda se orienta a problemas de Java en general, cuando sólo se utiliza el título en la clasificación, la baja performance puede deberse a que el nombre de las clases relacionadas no suele aparecer siempre en el título de los mismos; aunque sí suele aparecer al explicar el problema en la pregunta principal. En cuanto a la búsqueda considerando el hilo completo, la baja performance puede deberse a que las discusiones son largas y se pierde precisión al aumentar el contenido de las mismas (dan ejemplos o piden aclaraciones pero no aportan a la solución).

5 Análisis comparativo de los resultados

En la Figura 6 se contrastan los resultados agrupados según la cantidad de información en los hilos de discusión, considerando el agregado de sinónimos según lo establecido en la hipótesis B.2.

Table 3. F-Measure para cada corte de las pruebas realizadas

	2	3	4	5
O_aF_1	0.509	0.522	0.522	0.522
O_aF_2	0.573	0.581	0.568	0.572
O_aF_3	0.573	0.581	0.568	0.567
O_bF_1	0.4	0.39	0.402	0.401
O_bF_2	0.494	0.456	0.429	0.397
O_bF_3	0.433	0.408	0.394	0.367
O_cF_1	0.218	0.212	0.199	0.176
O_cF_2	0.167	0.183	0.166	0.54
O_cF_3	0.15	0.162	0.147	0.154

Al realizar una comparativa entre el primer caso de estudio (sin incorporar sinónimos) y el segundo caso de estudio (incorporando snónimos de adjetivos y adverbios), se observa que utilizar sinónimos mejora la performance respecto a considerar el texto original de los hilos cuando se tiene en cuenta el título y la pregunta principal (caso F_2), pero contrariamente a lo planteado, al utilizar el hilo completo al agregar sinónimos disminuye la performance (caso F_3). Esto puede deberse a que agregar todos los sinónimos de adjetivos y adverbios podría ocasionar ruido ya que los cálculos que realiza Lucene se realizan sobre el total de palabras de la consulta, lo cual no parece verse afectado en los casos F_2 y F_1 donde la cantidad de palabras es considerablemente más baja que en el hilo completo. Por otro lado, la incorporación de sinónimos al considerar título y pregunta (caso F_2) principal consigue la mejor performance de todas las combinaciones posibles, lo que apoyaría la hipótesis B.2 planteada.

Respecto a las amenazas a la validez del experimento, se debe mencionar que, los documentos Oracle utilizados para clasicar los hilos en estos casos de

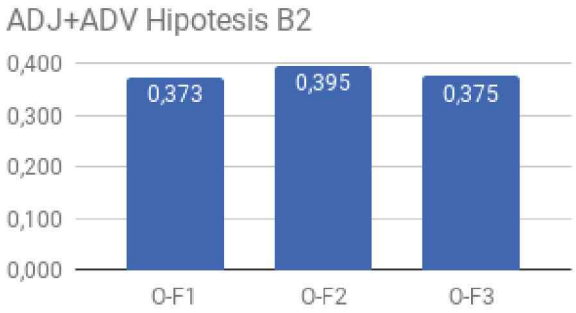


Fig. 5. Análisis de performance agrupado según cantidad de información en hilos de discusión (Hipótesis B2)

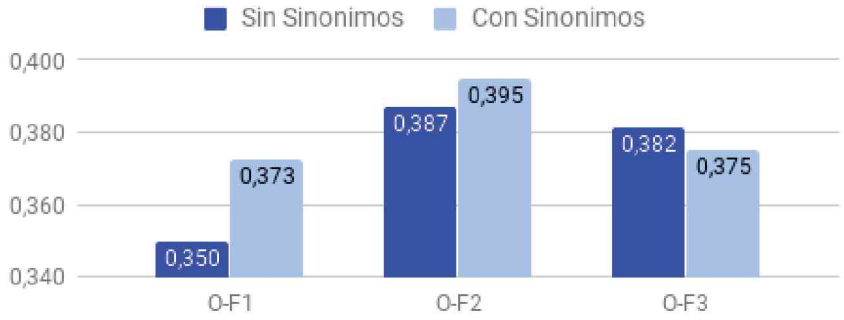


Fig. 6. Comparación de la performance al utilizar sinonimos de adjetivos y adverbios

estudio se restringieron a la versión Java 1.5, por lo que podría producirse alguna diferencia en los resultados al usar documentos de otra versión del lenguaje (por ejemplo por la inclusión o eliminación de un método de una versión a otra). Además, nuevas pruebas deben realizarse para evaluar estas hipótesis con un corpus de hilos mayor, recuperados de distintos foros.

6 Conclusiones y trabajo futuro

En este trabajo se presenta una estrategia para clasificar hilos de discusión recuperados de un foro de discusión técnico respecto a un conjunto de documentos de referencia, considerando la cantidad de información tomada de cada tipo de documento y la inclusión de sinónimos para lograr una mejor performance. Particularmente, en este artículo se presenta un caso de estudio con hilos recuperados del foro Stack Overflow, respecto al conjunto de documentos de especificaciones Oracle de las clases Java. La clasificación se realizó considerando las distintas secciones de ambos tipos de documentos y el agregado de sinónimos de los modificadores de sustantivos (adjetivos y adverbios) en los hilos de discusión.

De acuerdo a los resultados obtenidos con un corpus de 50 hilos de discusión, la clasificación tiene su mejor performance considerando el título y la pregunta principal del hilo e incluyendo sinónimos de los modificadores de sustantivos (adjetivos y adverbios).

Dado que este resultado proviene de un conjunto de hilos restringido, nuestro trabajo a futuro se enfocará en replicar estos experimentos con mayor cantidad de hilos así como con otras técnicas durante la fase de la recuperación de información, para asegurar la generalidad de estos resultados. Además, se pretende realizar nuevos experimentos controlados teniendo en cuenta las diferentes categorías gramaticales, utilizándolas individualmente y haciendo combinaciones de las mismas.

Posteriormente se plantea realizar resúmenes sobre los hilos de discusión, es decir, tener en cuenta las respuesta mas votadas, utilizar solo las respuestas

más relacionadas a la pregunta del usuario, etc., usando otras técnicas como la Similitud del Coseno, en lugar del hilo completo.

Finalmente, en este trabajo, las medidas utilizadas se aplican sobre un conjunto de documentos relacionados sin tener en cuenta su orden de relevancia. A futuro se planea ampliar el análisis a otros tipos de métricas que consideren el *ranking* de documentos relevantes retornado por Lucene.

Agradecimientos

Este trabajo está parcialmente soportado por el subproyecto "*Reuso de Conocimiento en Foros de Discusión - Parte II*", correspondiente al Programa de Investigación 04/F009 "*Desarrollo Orientado a Reuso - Parte II*"de la Universidad Nacional del Comahue (Neuquén, Argentina).

References

1. G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun, "Finding question-answer pairs from online forums," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, (New York, NY, USA), pp. 467–474, ACM, 2008.
2. S. Gottipati, D. Lo, and J. Jiang, "Finding relevant answers in software forums," in *26th IEEE/ACM International Conference on Automated Software Engineering (ASE 2011), Lawrence, KS, USA, November 6-10, 2011*, pp. 323–332, 2011.
3. S. Bhatia and P. Mitra, "Adopting inference networks for online thread retrieval.," in *AAAI*, vol. 10, pp. 1300–1305, 2010.
4. Aranda, Gabriela, Martínez Carod, Nadina, Roger, Sandra, Faraci, Pamela, and Cechich, Alejandra, "Una herramienta para el análisis de hilos de discusión técnicos," in *CACIC 2014, XX Congreso Argentino de Ciencias de la Computación*, (San Justo, Argentina), pp. 803 – 812, Oct. 2014.
5. V. Zoratto, G. N. Aranda, S. Roger, and A. Cechich, "Análisis de estrategias para clasificar contenidos en foros de discusión: Un caso de estudio," in *Simposio Argentino de Ingeniería de Software (ASSE 2015)-JAIIO 44*, (Rosario), pp. p. 176–190, SADIO, 2015.
6. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.
7. C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.
8. V. Zoratto, G. N. Aranda, S. Roger, and A. Cechich, "Analyzing discussion forums threads about java programming language usage.," *Electronic Journal of Informatics and Operations Research*, vol. 15, no. 1, 2016.
9. M. Nicoletti, S. Schiaffino, and D. Godoy, "Mining interests for user profiling in electronic conversations," *Expert Syst. Appl.*, vol. 40, pp. 638–645, Feb. 2013.
10. R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.